View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by idUS. Depósito de Investigación Universidad de Sevilla

REGRESANDO AL PASADO: OTRA FORMA DE INTRODUCIR LA REGRESIÓN ESTADÍSTICA

José Antonio Camúñez Ruiz
Mª Dolores Pérez Hidalgo
Francisco Javier Ortega Irizo
Departamento de Economía Aplicada I
Facultad de Ciencias Económicas y Empresariales. Universidad de Sevilla

RESUMEN

Para iniciar al estudiante en el concepto de Regresión Lineal, y en el de Correlación asociado, en un curso introductorio de Estadística Descriptiva se propone el uso de problemas históricos relacionados con la materia, en particular, de los problemas abordados por Galton a finales del siglo XIX, asociados a la herencia genética. Además de trasladar al alumno al contexto histórico en el que surgieron, dándoles a conocer nombres de científicos pioneros, se procede con la resolución de esos problemas de manera intuitiva, visualizándolos gráficamente, probando con diferentes medidas de la estadística descriptiva ya conocidas por ellos, con ensayo y error, con el apoyo de calculadora y hoja de cálculo. Para el final dejamos la formalización de la teoría. La experiencia propone, pues, un recorrido inverso al que suele ser habitual. Los resultados de la misma, valorados por tres vías, resolución de ejercicios prácticos, encuesta a los alumnos y nota del examen, muestran que el proceso ha sido positivo para el objetivo último del aprendizaje.

Palabras clave: Regresión, correlación, Galton, Pearson, hoja de cálculo.

ABSTRACT

In an introductory course in Descriptive Statistics, to initiate students into the concept of linear regression and the associated correlation, we suggest the use of historical problems related to the subject, specially, the problems addressed by Galton in the late of the 19th century, associated with genetic inheritance. Besides moving the student to the historical context in which they emerged and announcing names of leading scientists, we proceed with the resolution of these problems in an intuitive way, graphically visualizing them, trying with different measures of descriptive statistics already known by students, through trial and mistake, with the support of calculator and spreadsheet. Finally, we formalize the theory. The experience suggests, therefore, a reverse path than currently. The results, measured in three ways (solving exercises, surveying students and qualifying exam) show that the process has been positive for the ultimate goal of learning.

Keywords: Regression, correlation, Galton, Pearson, spreadsheet.

1. INTRODUCCIÓN. ANTECEDENTES

El análisis de regresión es uno de los pilares fundamentales en los programas de aquellas asignaturas que inician al alumno en el aprendizaje de las técnicas estadísticas. Es bien conocido, por otra parte, que dichas técnicas se convierten en valiosos instrumentos para el desarrollo profesional de los titulados en Administración y Dirección de Empresas. Las asociaciones del tipo causa-efecto, las relaciones entre variables, los niveles de intensidad de dichas relaciones, las predicciones de comportamientos a partir de unos indicios, todo ello constituye un conjunto de ideas que, a nivel estadístico, se enmarca en el amplio concepto de "regresión". En los cursos de iniciación, el concepto se introduce de manera descriptiva, manejando datos observados. Después, una vez que el alumno se inicia en el cálculo de probabilidades y, por tanto, en los modelos probabilísticos teóricos, de nuevo la regresión se incorpora a los programas de las asignaturas, pero ahora entre variables aleatorias. En este caso, desde luego, los conceptos ya adquiridos a nivel descriptivo son fundamentales. Por último, en la disciplina conocida como Econometría, que forma parte del currículo de la mencionada titulación en cursos superiores, la regresión es la base sobre la que se sustenta la mayor parte de las técnicas de análisis econométrico. Somos conscientes, por tanto, de la trascendencia que tiene este capítulo y nos preocupa que la absorción de conocimientos e ideas sea completa y, además, que resulte atractiva.

Tradicionalmente, este capítulo ha sido impartido siguiendo esta dirección: en primer lugar se introduce el coeficiente de correlación lineal, o sea, el concepto de asociación lineal entre variables (con su fórmula asociada) y, a continuación, la recta de regresión con el cálculo teórico de coeficientes y medidas de bondad de ajuste. Todo ello acompañado de importantes y, a veces, tediosos desarrollos algebraicos (aunque necesarios), para la justificación de las diferentes fórmulas. La conexión entre correlación y regresión queda oscurecida al producirse más a nivel algebraico que a nivel de ideas. El tema se completa con una serie de ejercicios prácticos y problemas, primero desarrollados por el profesor, y después propuestos, donde el alumno ha de mostrar los conocimientos y habilidades adquiridos durante las explicaciones previas, tanto teóricas como prácticas, del profesor. Todo el proceso de aprendizaje lleva una dirección vertical, de arriba abajo, del profesor al alumno, sin apenas reflexión y maduración sobre las nuevas e importantes ideas. La inmensa mayoría de los manuales publicados sobre la materia siguen ese camino. Nuestra experiencia como profesores (hasta 20 años de docencia) nos muestra un panorama algo sombrío e insatisfactorio, por quedar este tópico un poco deslavazado en ese océano inmenso de la estadística.

2. OBJETIVOS

Lo descrito nos ha hecho plantear, desde hace algún tiempo, la búsqueda de alternativas docentes. Una de ellas sería (y es la que aquí hemos desarrollado) la misma que sirvió a Galton para introducirse en estos conceptos a finales del siglo XIX. Galton, primo de Darwin, y reconocido científico de ese siglo por derecho propio, ha sido con frecuencia criticado por su apuesta por la eugenesia. Por otra parte, hay quien opina que

la fama perdurable en el tiempo de su primo ha ensombrecido de forma injusta las importantes aportaciones científicas con las que Galton contribuyó al campo de la biología, la psicología y la estadística aplicada. Su pasión por la genética y, en particular, por los problemas de herencia, es lo que le llevó a pensar en métodos de cálculo como la regresión y la correlación. Así, las reflexiones que le llevan a este campo comienzan con un complicado (entonces) problema de herencia: la comprensión de la fuerza con la que las características de una generación de seres vivos se manifiesta en la siguiente. Inicialmente, Galton se aproxima a este problema examinando características de semillas de guisante. Elige el guisante porque esta especie puede autofertilizarse: las plantas hijas manifiestan variaciones genéticas de las plantas madres sin la contribución de un segundo progenitor. De esta forma, Galton pospone el problema de calcular estadísticamente las contribuciones genéticas de varias fuentes. La primera idea de Galton sobre la regresión nace de un gráfico, un diagrama bidimensional, en el que se representaba los tamaños de los guisantes hijos frente a los de los guisantes padres. Galton se dio cuenta que el diámetro mediano de las semillas hijos para un diámetro concreto de la semilla padre describe, aproximadamente, una línea recta con pendiente positiva y menor que 1. Este autor usa la representación de sus datos para ilustrar los fundamentos básicos de lo que los estadísticos seguimos llamando regresión. A partir de aquí, con los errores propios de cualquier proceso incipiente de investigación, Galton empieza a construir toda una teoría que, matemáticamente, fue formalizada posteriormente por uno de sus discípulos, Pearson.

Entonces, el objetivo docente de la experiencia desarrollada ha sido una mezcla entre "aprendizaje basado en problemas" y "nacimiento y desarrollo histórico". Pragmatismo e historia. La modelización matemática ha sido a posteriori. Invertimos el orden, de la práctica a la teoría, de los alumnos al profesor: los problemas motivan, los alumnos piensan y proponen soluciones, y el profesor supervisa y orienta. Con ello intentamos mejorar la comprensión de los fundamentos y animar el interés del estudiante al mostrarse los diversos problemas con los que Galton, y otros investigadores tempranos se enfrentaron y solucionaron cuando ellos iniciaron las técnicas que tan extensamente son usadas hoy.

La experiencia se ha desarrollado en un grupo de unos 80 alumnos de la doble licenciatura de ADE y Derecho en la asignatura Estadística I que constituye un curso introductorio a la estadística y en el que la mayor parte de su contenido está relacionado con técnicas y métodos de Estadística Descriptiva.

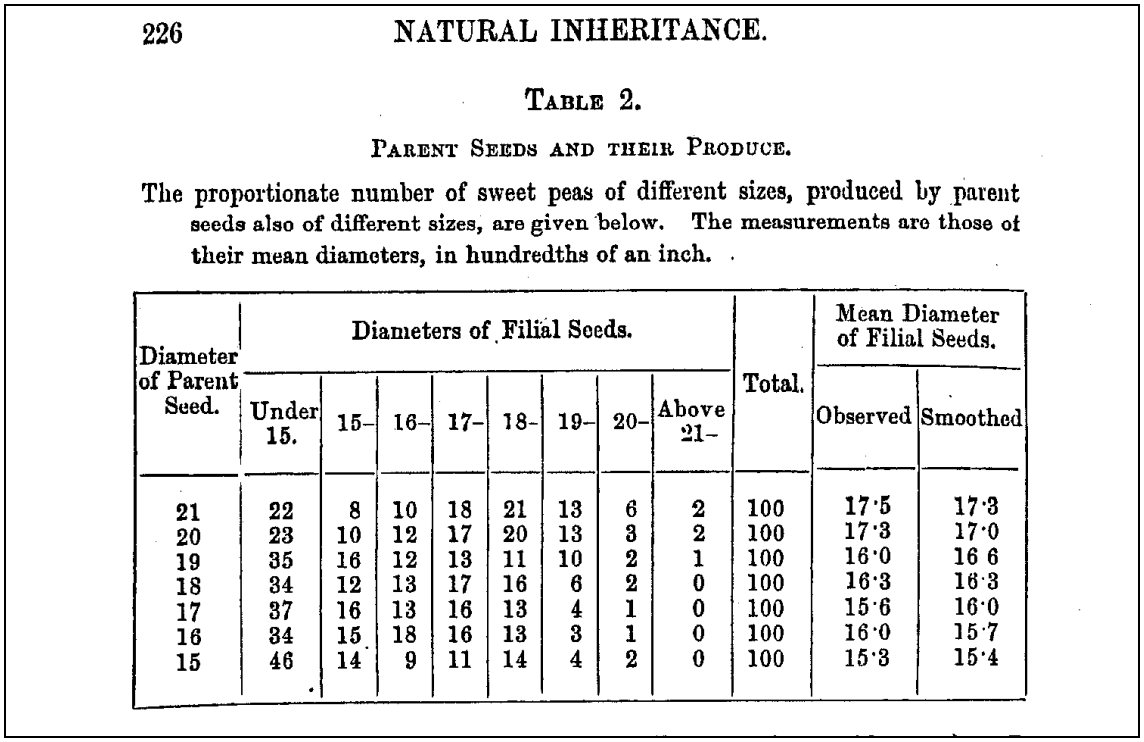
3. METODOLOGÍA

El camino seguido, entonces, ha ido en la siguiente dirección:

1. En primer lugar, algunos aspectos biográficos de Sir Francis Galton son reseñados. Se remite a los alumnos a la página Web reconocida como oficial, sobre la vida y obra de este autor: http://galton.org/. Esta circunstancia permite un recuento histórico de los aspectos de progreso científico y matemático a finales del siglo XIX y principios del XX. En este contexto, se ha descrito el bagaje científico con el que Galton se enfrenta al problema y se ha explicado cómo sus carencias matemáticas

limitaron en un principio el desarrollo analítico de la regresión. Los alumnos acceden, además, a los problemas más acuciantes para la ciencia a finales del siglo XIX, siendo uno de ellos el de la herencia genética.

2. Empleo en el aula de algunos de los ejemplos históricos, con los mismos datos que estos pioneros. El primer conjunto de datos que se le ofreció al alumnado es el que Galton trabajó en su famosa obra *Natural Inheritance* (1894) En sus cuatro volúmenes biográficos, Pearson describe la génesis del descubrimiento de la pendiente de regresión (Pearson 1930). En 1875 Galton distribuyó paquetes de semillas de guisantes entre siete amigos; cada uno de ellos recibió semillas de diámetro uniforme (ver también Galton 1894), pero había sustanciales diferencias entre los diferentes paquetes. Los amigos de Galton recolectaron las semillas de la nueva generación y se la devolvieron a él. Las medidas de los diámetros de esta segunda generación cruzadas con las de los progenitores la recoge el autor en la siguiente tabla, siendo ésta la primera que se les facilita a los alumnos:



226 NATURAL INHERITANCE.

TABLE 2.

PARENT SEEDS AND THEIR PRODUCE.

The proportionate number of sweet peas of different sizes, produced by parent seeds also of different sizes, are given below. The measurements are those of their mean diameters, in hundredths of an inch.

Diameter of Parent Seed.	Diameters of Filial Seeds.								Total.	Mean Diameter of Filial Seeds.	
	Under 15.	15–	16–	17–	18–	19–	20–	Above 21–		Observed	Smoothed
21	22	8	10	18	21	13	6	2	100	17·5	17·3
20	23	10	12	17	20	13	3	2	100	17·3	17·0
19	35	16	12	13	11	10	2	1	100	16·0	16 6
18	34	12	13	17	16	6	2	0	100	16·3	16·3
17	37	16	13	16	13	4	1	0	100	15·6	16·0
16	34	15	18	16	13	3	1	0	100	16·0	15·7
15	46	14	9	11	14	4	2	0	100	15·3	15·4

Gráfico 1: Imagen de la Tabla publicada por Galton sobre el diámetro de semillas de guisantes padres versus semillas de guisantes hijos.

Con los datos de esa tabla, se propone la construcción de una base de datos tipo Excel, en la que se especifiquen tres columnas: las dos variables a relacionar, diámetro de la semilla padre, los de las semillas hijos, y como tercera columna la frecuencia absoluta, o sea, el número de veces que una determinada pareja se repite. Dado que los datos correspondientes a las semillas hijos los presenta Galton

agrupados en intervalos, se selecciona una "marca de clase" para cada uno. Un ejemplo de construcción de esa base de datos es la tabla que sigue:

Diámetro semilla padre	Diámetro semilla hijo	Frecuencia		Diámetro semilla padre	Diámetro semilla hijo	Frecuencia
21	14,5	22		18	17,5	17
21	15,5	8		18	18,5	16
21	16,5	10		18	19,5	6
21	17,5	18		18	20,5	2
21	18,5	21		17	13,5	37
21	19,5	13		17	14,5	16
21	20,5	6		17	15,5	13
21	22,5	2		17	16,5	16
20	14,5	23		17	17,5	13
20	15,5	10		17	18,5	4
20	16,5	12		17	19,5	1
20	17,5	17		16	14,5	34
20	18,5	20		16	15,5	15
20	19,5	13		16	16,5	18
20	20,5	3		16	17,5	16
20	22,5	2		16	18,5	13
19	14,5	35		16	19,5	3
19	15,5	16		16	20,5	1
19	16,5	12		15	13,5	46
19	17,5	13		15	14,5	14
19	18,5	11		15	15,5	9
19	19,5	10		15	16,5	11
19	20,5	2		15	17,5	14
19	22,5	1		15	18,5	4
18	14,5	34		15	19,5	2
18	15,5	12				
18	16,5	13				

Tabla 1: Datos en excel de la Tabla publicada por Galton sobre el diámetro de semillas de guisantes padres versus semillas de guisantes hijos.

En una gráfica, Galton presentó los diámetros de los guisantes padres frente a los de los hijos. Como se ha dicho, descubre como los diámetros medianos de las semillas hijos para cada diámetro concreto de la semilla progenitor describe, aproximadamente, una línea recta con pendiente positiva y menor que 1. Así, de forma natural, encontró una primera recta de regresión y, también, una variabilidad constante para todas las series de un segundo carácter, para un carácter dado del primero. Quizás, el estudio de este caso simple tan especial fue lo mejor para el progreso del cálculo correlacional, dada la facilidad para su comprensión por parte de un principiante. Por tanto, siguiendo ese proceso de conceptualización, se

propone en primer lugar el uso de métodos gráficos. Mostramos a los alumnos la representación gráfica del diagrama de dispersión asociado a estos datos. Excel permite esta representación. En el aula disponemos de ordenador para el profesor, con proyección en pantalla y, a su vez, los alumnos acuden a clase con ordenador portátil en su mayoría. Ahora bien, con el fin de facilitar la introducción teórica del concepto, aconsejamos que procedan a la tipificación de ambas variables, quedando así las dos centradas en el origen de coordenadas (las dos medias se transforman en (0,0)) y ambas con igual variabilidad (sus desviaciones típicas se hacen igual a 1), de manera que las diferencias en sus magnitudes o escalas no interfieran (enmascarando o engordando) en el análisis de la posible relación entre ambas. Aunque advertimos a los alumnos que, en el caso que nos ocupa, las dos variables son muy parecidas en cuanto a centralidad y variabilidad por lo que, quizás, no hubiese sido necesario sus tipificaciones. Con las variables tipificadas representamos y, así, visualizamos la relación, invitando a los alumnos a la búsqueda de funciones sencillas (rectas) que sean capaces de reflejar lo mejor posible lo que el diagrama de dispersión transmite. Acompañamos cada punto del diagrama de un valor numérico que coincide con la frecuencia absoluta correspondiente. De alguna forma, ese valor nos informa del peso que debe tener el punto asociado en el análisis de esa relación.

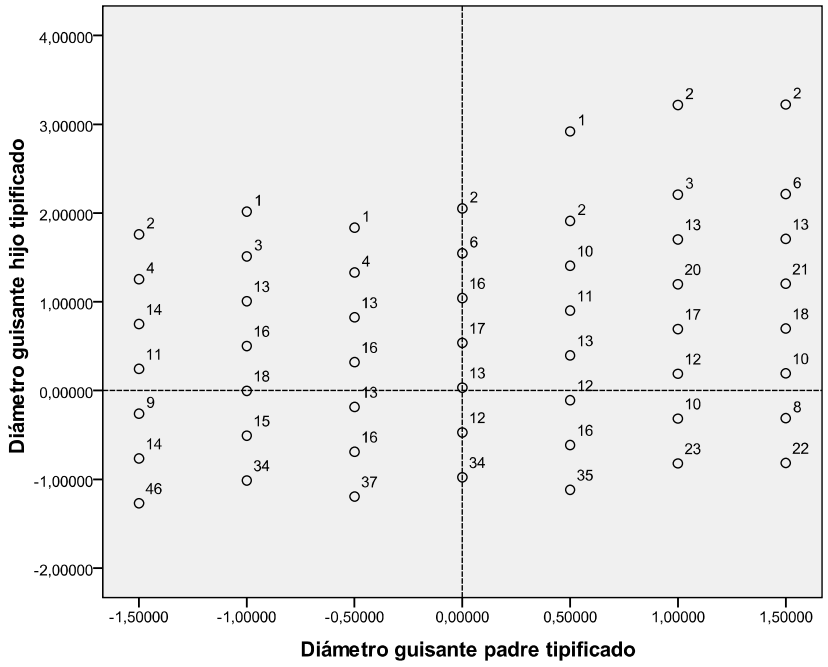


Gráfico 2: Diagrama de dispersión de los datos de Galton tipificados junto con las frecuencias absolutas de las correspondientes parejas.

El trazo aproximado de una recta intercalada entre los puntos que intente aproximarse lo máximo posible a todos ellos, que pase por el origen de coordenadas al ser éste el "centro" del diagrama y que tenga en cuenta el "peso" de cada punto en la gráfica es lo que Galton intentó como primera aproximación y es lo que se plantea

ante la visión de los datos en el plano. Una vez que cada alumno ha trazado su propia recta se le invita al cálculo de la pendiente de la misma, cálculo fácil por otra parte al tratarse de un cociente entre distancia opuesta y contigua. En ese contexto el alumno se da cuenta de dos detalles: la recta es creciente (los puntos situados en el tercer cuadrante tienen mucho peso), pero con pendiente suave (desde luego bastante menor que 1 que correspondería a la bisectriz del 1º y 3º cuadrante). El primero nos informa de una relación directa entre ambas variables: en general, a diámetros pequeños del padre corresponden pequeños del hijos y, por el otro extremo, a diámetros grandes del progenitor, diámetros grandes de la semilla hijo. El segundo nos hace pensar que grandes variaciones en los diámetros progenitores se traducen en variaciones más pequeñas en la segunda generación, o sea, valores más próximos al centro en esta segunda (lo que Galton llamó "regresión a la media", que dio origen al término que nomina toda esta teoría).

Una segunda propuesta gráfica es planteada. Los alumnos conocen ya la construcción de los diagramas de cajas. En este segundo gráfico la mediana hace el papel que hacía la media en el anterior. Informamos que los primeros intentos de Galton de construcción de sus regresiones fueron usando la mediana que, aunque a él le resultaba más intuitiva, presenta dificultades algebraicas que impiden los desarrollos de cálculo asociados.

De todas formas, esta visualización corrobora la idea que ya se estaba formando. Una recta que se intercale pasaría por el origen de coordenadas, tendría pendiente positiva, pero ésta sería menor que uno:

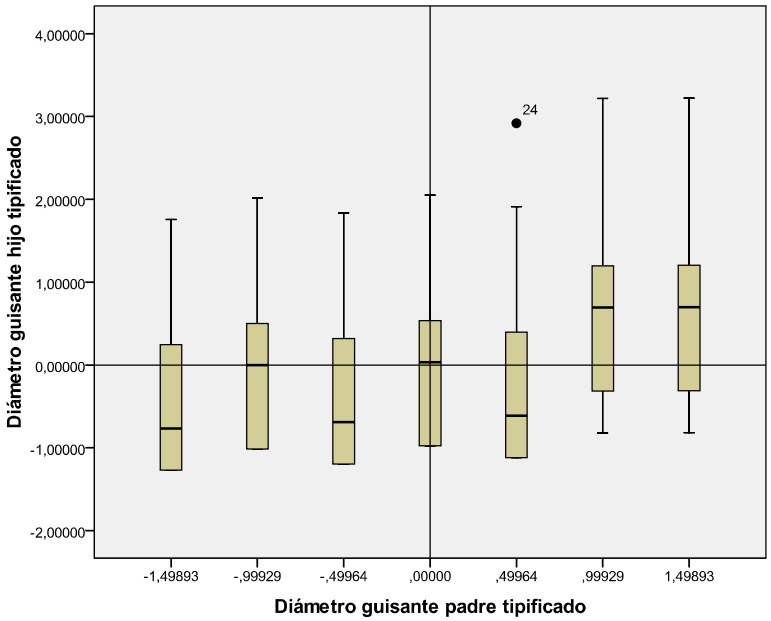


Gráfico 3: Diagrama de cajas de los datos de Galton tipificados.

3. Mediante calculadora científica, primeramente, y hoja de cálculo Excel, después, los alumnos estiman la "pendiente" que, de alguna forma, definirá la posible relación. La facilidad de los instrumentos permiten repetir diferentes cálculos e, incluso,

intentar diferentes medidas, hasta encontrar las mejores aproximaciones. Así pues, aplicamos métodos de ensayo y error. En las aproximaciones realizadas por los estudiantes en esta ocasión encontramos pendientes comprendidas entre 0'30 y 0'50. A dicha pendiente Galton la llamó *r* (de regresión).

4. Es el momento de la formalización de la idea. La intuición nos hace pensar que lo planteado es un problema de optimización: obtener la mejor recta que represente la nube de puntos, o sea, que minimice la suma de distancias (al cuadrado) de los puntos del diagrama de dispersión a la recta. El problema recuerda uno ya resuelto al trabajar con variables unidimensionales: minimización de las distancias (al cuadrado) de los valores de la distribución a un valor central o, dicho de otra forma, intento de sustitución de los valores de esa distribución por un único valor que los represente, solución que ofrece el Teorema de König (conocido ya por los alumnos a esta altura del desarrollo de la disciplina) y que lleva directamente a la **media aritmética** como valor central representante y óptimo en el sentido de las distancias. Sólo queda dar el salto al caso bidimensional. En lugar de una variable, tenemos dos. Generalizando, la solución óptima es de nuevo una media aritmética, siendo en este caso la de la variable resultante al construir los productos cruzados de las variables originales. Por tanto, la solución, la pendiente buscada, es la media de los productos cruzados entre ambas variables o, lo que Pearson llamó, "momento-producto". Podemos escribir, entonces, $r = \overline{XY}$, y la ecuación de la recta que "ajusta" lo mejor posible la nube de puntos viene dada por $y = rx$, donde y representa la variable "diámetro de guisantes hijos tipificados", o sea, la variable efecto, o variable a explicar, o dependiente, mientras que x es la de los "diámetros de los guisantes padres tipificados", variable causa, o explicativa, o independiente. La recta, como se esperaba, pasa por el origen de coordenadas. Para los datos de Galton tipificados obtenemos $r = 0'346$. Entonces, la recta ajustada queda representada entre la nube de puntos:

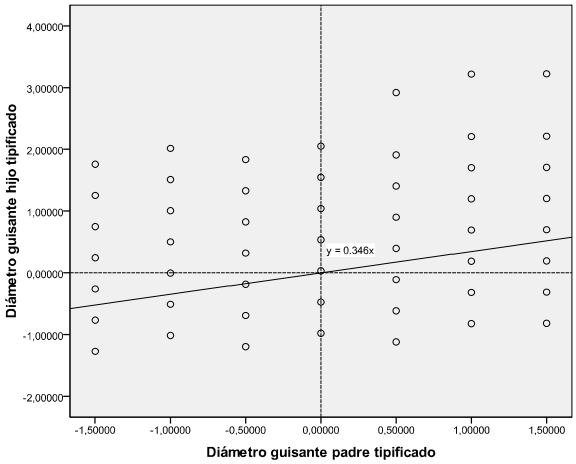


Gráfico 4: Diagrama de dispersión de los datos de Galton tipificados junto con la recta ajustada usando como pendiente el momento-producto.

5. Con tres ejemplos sencillos, con pocos datos, se proponen casos en los que las variabilidades coinciden o son distintas. Seguimos usando datos centrados en el origen para evitar aún la introducción de la constante en la recta.

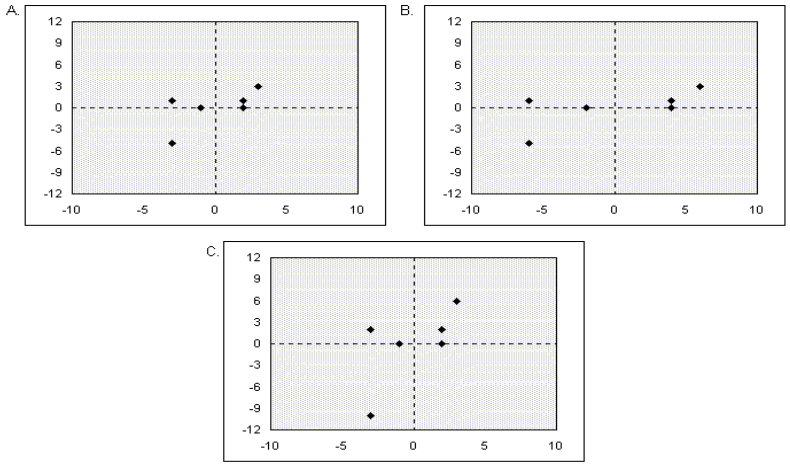


Gráfico 5: Diagrama de dispersión de datos de tres ejemplos sencillos.

El objetivo es, a partir de la pendiente inicialmente construida para las dos variables con igual dispersión, derivar las pendientes resultantes en los casos en que la dispersión de *x* es mayor que la de *y*, cuando ambas son iguales (como en el caso de los datos anteriores tipificados) y cuando la variabilidad de *x* es menor que la de *y*. El alumno capta cómo esas variabilidades, medidas a través de sus desviaciones típicas, van haciendo que la recta se incline (cuando $S_x > S_y$) o se empine (cuando $S_x < S_y$). Entonces, con diversas pruebas de cálculo y apoyándonos en la propia optimación anterior, llegamos a formular la siguiente ecuación: $y = r\left(S_y / S_x\right)x$.

Desde este instante, el alumno comienza a distinguir los dos parámetros fundamentales del análisis de regresión: *r*, al que llamamos coeficiente de correlación (grado de relación lineal entre las dos variables) y $b = r\dfrac{S_y}{S_x}$, la pendiente de la recta, donde en el cociente aparecen las desviaciones típicas de las dos variables relacionadas. Los tres ejemplos anteriores han sido preparados para que tengan el mismo valor de *r*, para que así, el que se inicia en este estudio capte que el cambio de las pendientes se debe, en estos casos, a la diferencia existente en la variabilidad de ambas. Conseguimos que el alumno sea consciente de la diferencia entre correlación y pendiente, y que acepte el cociente $\dfrac{S_y}{S_x}$ como factor corrector o igualador entre ambos términos. De paso, la proporción está servida: la pendiente es a la desviación típica de *y* como el coeficiente de correlación es a la de *x*, o la proporción entre pendiente y coeficiente de correlación es la misma que entre la

desviación típica de *y* y la de *x*: $\frac{b}{r}=\frac{S_y}{S_x}$. La devolución a las variables originales, deshaciendo las tipificaciones, hace aflorar de manera sencilla la constante o término independiente en la recta.

6. Mediante Excel calculamos valores ajustados y errores o residuos (diferencias entre observados y ajustados). Es fácil comprobar en la hoja de cálculo que los residuos suman cero, que el efecto compensatorio entre errores positivos y negativos es total con la recta ajustada. Pero los errores existen y pueden ser mayores o menores dependiendo del nivel de dispersión de la nube de puntos. La necesidad de una medida de bondad de ajuste surge. Ha de ser un agregado, una medida resumen de los errores. Como los errores pueden ser positivos y negativos, una medida cuadrática evita el posible efecto compensatorio del signo. Se propone entonces la suma de cuadrados de errores o residuos: SCR, que valdrá 0 cuando el ajuste sea perfecto y será tanto mayor cuanto más dispersión haya en el diagrama. Una forma de dar seguridad al trabajo es proponer al alumno el uso de otros valores para la pendiente de la recta, o sea, otros ajustes de la misma, y el correspondiente cálculo asociado de la SCR para mostrar que, en cualquier caso, ese agregado de errores siempre es mayor. La SCR es una medida absoluta de la bondad de ajuste. La misma, como es lógico, depende de la escala empleada para la variable dependiente. Por otra parte, podemos calcular la variabilidad de dicha variable, mediante su varianza, que representamos por VT (varianza total), también la variabilidad de los valores ajustados o explicados, VE (varianza explicada) y, por último, la variabilidad de los errores o residuos, que no es más que la SCR convertida en media al dividirla por el tamaño muestral y que representamos por VR (varianza residual). Los cálculos llevan al alumno a comprobar la siguiente relación intuitiva: $VT=VE+VR$, base fundamental de todo el cálculo de la regresión. La misma nos lleva a plantear una segunda medida de bondad de ajuste, en este caso relativa y, por tanto, útil para comparar con otros ajustes: "Proporción de varianza explicada". La llamamos Coeficiente de Determinación y su representación y cálculo quedan definidos en la siguiente igualdad $R^2=\frac{VE}{VT}$. La sorpresa del estudiante es mayúscula cuando comprueba que este coeficiente coincide con el cuadrado de *r* lo que, de paso, justifica la simbología empleada para representarlo.
7. El salto a la regresión múltiple será natural y, para el alumno, casi necesario cuando sea consciente de la necesidad de introducir más de un factor influyente en la variable objetivo. Ilustramos como Galton se dio cuenta, poco después de haber recogido y analizado sus datos sobre guisantes, que la generación previa a los padres inmediatos también puede influir en las característica individuales (Pearson, 1930). Señala que, incluso, ciertas características se saltan una o más generaciones, ocasionalmente; un hombre puede ser más parecido a su abuelo que a su padre, en ciertos aspectos. En un artículo de 1898 en la revista *Nature* (citado en Pearson, 1930), Galton publicó un ingenioso diagrama que particionaba un cuadrado unidad en sucesivos cuadrados más pequeños, donde cada uno representaba la cada vez más

disminuida influencia de las generaciones previas de los ancestros sobre el individuo actual. Galton dio con el germen de la idea de regresión múltiple. Una característica o variable puede ser influida no sólo por una única causa importante, sino por muchas causas de mayor y menor importancia. Algunas de estas causas, incluso, pueden superponerse entre ellas (esto es, las variables explicativas están correlacionados entre si). En publicaciones posteriores Galton listó algunas fórmulas matemáticas que recogían esta misma idea básica, pero nunca fue capaz de desarrollar un tratamiento matemático completo del asunto:

"Las matemáticas algo complicadas de la correlación múltiple, con sus repetidas apelaciones a las nociones geométricas de hiperespacio, le dejaron una habitación cerrada." (Pearson 1930, p. 21)

Sin embargo, la conceptualización de Galton de las múltiples influencias de los antepasados sobre las características del individuo del presente era completamente paralela a la concepción moderna de regresión múltiple. Como con la regresión lineal simple y el coeficiente de correlación, Galton puso el trabajo preliminar imaginativo que Pearson más tarde desarrolla con un tratamiento matemático riguroso. El trabajo subsiguiente de Pearson incluyó el posterior desarrollo de la regresión múltiple así como el progreso innovador en otros estadísticos. Entonces, usando la formulación (aunque actualizada) y ejemplos de Pearson introducimos al alumno en la regresión múltiple.

4. RESULTADOS

Concluida la experiencia, los profesores que intervenimos en la misma disponemos de tres instrumentos para evaluar su resultado: los trabajos desarrollados por los alumnos durante las clases, la encuesta a los alumnos y la calificación del examen sobre esta materia.

Otros ejemplos prácticos fueron propuestos, todos extraídos de la historia de esta disciplina. Algunos desarrollados en clase, con asistencia y supervisión de los profesores. Se pedía a los alumnos, además de los cálculos adecuados, la elaboración de un informe concluyente sobre lo extraído de esos cálculos. De esta forma, aprovechamos para iniciar a los mismos en la redacción de ese tipo de informe. Los trabajos, redactados en Word y remitidos vía e-mail fueron calificados en una escala de 0 a 10. Las calificaciones de estos trabajos arrojaron los siguientes resultados:

- La calificación media fue 7'71 con una desviación típica de 2'03.
- El percentil 50, o mediana fue 8'50, mientras que el percentil 75 toma el valor de 9'1, lo que nos da una idea de mayoría de calificaciones entre notable y sobresaliente.
- Un histograma de estas calificaciones es el que aparece en el Grafico 6.

La encuesta a los alumnos se desarrolló al final de la experiencia. Se les presentó una serie de afirmaciones sobre las que los mismos manifestaban desde su "total desacuerdo" hasta su "total acuerdo", en una escala tipo Likert de cinco categorías, que recorren el camino señalado. Al final de la encuesta se les pedía que hiciesen una valoración global de la experiencia mediante una calificación, en una escala de 0 a 10.

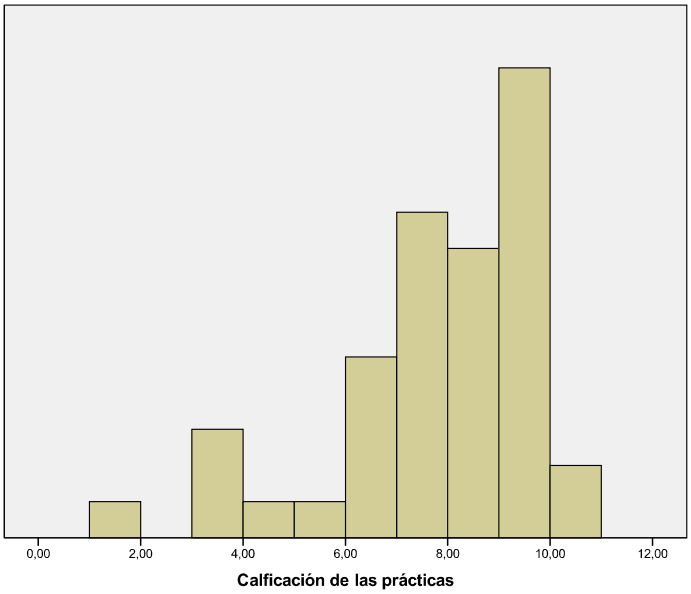


Gráfico 6: Histograma de las calificaciones de las prácticas.

Las siguientes tablas muestran resultados porcentuales de algunas de las afirmaciones planteadas en la encuesta.

Afirmación: *El contexto histórico me ha ayudado a entender el por qué de la regresión.*

Posibles Respuestas	Porcentaje
Ni de acuerdo ni en desacuerdo	2,9
De acuerdo	57,1
Totalmente de acuerdo	40,0

Afirmación: *La experiencia desarrollada me ha servido para soltarme en el manejo de Excel.*

Posibles Respuestas	Porcentaje
En desacuerdo	5,7
Ni de acuerdo ni en desacuerdo	25,7
De acuerdo	57,1
Totalmente de acuerdo	11,4

Afirmación: *El aprendizaje de la Estadística queda más motivado usando contextos históricos.*

Posibles Respuestas	Porcentaje
Ni de acuerdo ni en desacuerdo	11,4
De acuerdo	60,0
Totalmente de acuerdo	28,6

En cuanto a la calificación que los estudiantes dan a esta experiencia la resumimos en los estadísticos que aparecen en la siguiente tabla.

Calificación de la experiencia

Media		7,7
Mediana		8,0
Moda		8,0
Desviación típica		0,7
Percentiles	25	7,0
	50	8,0
	75	8,0

Por ultimo, mostramos lo que, quizás, más nos interesa como docentes: conocer si la experiencia desarrollada ha contribuido positivamente en el aprendizaje del alumno. La forma en la que pretendemos evaluar es el examen escrito, similar en cuanto a su estructura y contenidos al de anteriores cursos. Los estadísticos más importantes relacionados con la nota del examen correspondiente a esta materia son los siguientes:

Notas del examen

Media		6,04
Moda		10,00
Desviación típica		2,39
Percentiles	25	3,90
	33	5,00
	50	6,10
	61	7,00
	75	7,70

Por tanto, un 67% de los alumnos superaron la materia, siendo casi un 40% los calificados con notable o sobresaliente. Hemos tomado las calificaciones correspondientes a este mismo grupo, pero del curso anterior, y mediante una prueba *t* para muestras independientes, comparamos las notas para esos dos años consecutivos. Mostramos resultados:

Curso	Media	Desviación típica
2008-2009	3,93	2,33
2009-2010	6,04	2,39

Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias				
F	p-valor	t	gl	p-valor (prueba bilateral)	Diferencia de medias	Error típ. de la diferencia
0,000	0,989	-4,131	84	**0,000**	-2,10	0,51

Asumiendo igual variabilidad en las calificaciones de uno y otro curso (lo que confirma la prueba de Levene), la diferencia entre las notas medias, de algo más de 2 puntos, en una escala de 0 a 10, a favor de las notas del último año (cuando se llevó a cabo la experiencia) resulta estadísticamente significativa (p=0'000) según esta prueba *t* de Student. La comparación de los diagramas de cajas asociados visualiza estos resultados:

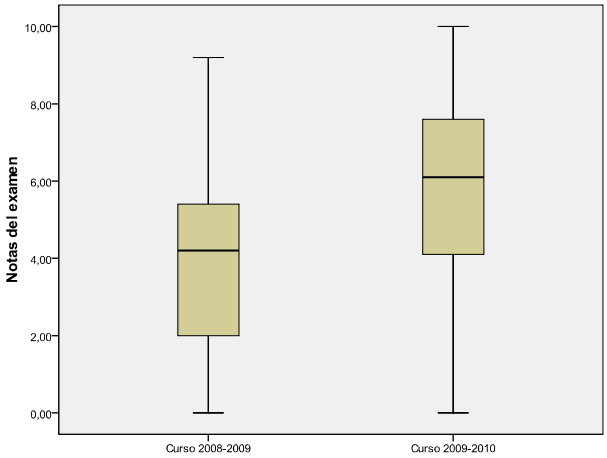


Gráfico 7: Diagramas de cajas de las notas de dos cursos consecutivos

No somos tan osados como para pensar que la diferencia en las notas medias de ambos cursos sea debida, exclusivamente, al desarrollo o no de la experiencia comentada. Somos muy conscientes, y nuestra experiencia como docentes así nos lo dice, que en cada grupo y en cada curso intervienen muchos factores, algunos conocidos por el profesor y otros no, que influyen en las notas del examen. Por tanto, nos resulta difícil valorar, cómo ha pesado el desarrollo de la experiencia en esa diferencia de notas. Pensamos, que los resultados presentados en este apartado sí pueden dar una visión, aunque sea aproximada, que podría ser válida para efectuar una valoración positiva a esta propuesta.

5. CONCLUSIONES

La historia del progreso científico es un buen argumento didáctico. Enfrentar al estudiante a los mismos problemas con los que científicos pioneros establecieron las bases de nuevas teorías es una forma más de motivar el aprendizaje. Si a ello unimos un

proceso de ensayo y error, intuitivo, de diferentes aproximaciones a la solución del problema, dadas las posibilidades que nos permite una hoja de cálculo, podemos asegurar el refuerzo y comprensión de la materia que se imparte. El formalismo teórico a posteriori. Resumiendo: intuición y cálculo, observación y reflexión. Recorrido de un camino que ya científicos de importante talla lo siguieron en su momento.

6. BIBLIOGRAFÍA

DUKE, J. D. (1978). Tables to Help Students Grasp Size Differences in Simple Correlations. **Teaching of Psychology, 5**, 219-221.

FITZPATRICK, P. J. (1960). Leading British Statisticians of the Nineteenth Century. **Journal of the American Statistical Association, 55**, 38-70.

GALTON, F. (1894). **Natural Inheritance (5th ed.)**. New York, Macmillan and Company.

GOLDSTEIN, M. D., STRUBE, M. J. (1995). Understanding Correlations: Two Computer Exercises. **Teaching of Psychology, 22**, 205-206.

KARYLOWSKI, J. (1985). Regression Toward the Mean Effect: No Statistical Background Required. **Teaching of Psychology, 12,** 229-230.

PEARSON, E. S. (1938). **Mathematical Statistics and Data Analysis (2nd ed.)**. Belmont, CA: Duxbury.

PEARSON, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. **Philosophical Transactions of the Royal Society of London, 187**, 253-318.

PEARSON, K. (1922). **Francis Galton: A Centenary Appreciation.** Cambridge University Press.

PEARSON, K. (1930). **The Life, Letters and Labors of Francis Galton**. Cambridge University Press.

STANTON, J. M. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. **Journal of Statistics Education Vol 9, N. 3.**